International Journal of Computer Science and Engineering Communications
Volume.4, Issue.3 (2016): Page.1423-1427
www.scientistlink.com/ijcsec

A Reduction of Online Execution Time with Sparql Query Method on Mapreduce for Bigdata Applications

Suganya.C[1], Raji.V[2], Kumaresan.A[3], Nanduja.R[4]

[1,4]PG Scholar, [2]Assistant Professor, [3]HOD,

Department of Computer Science, SKP Engineering College, India

[1]suganyacmecse@gmail.com, [2]raji.vpm@gmail.com, [3]kummaresan@gmail.com, [4]nandhu.priya.cse@gmail.com

***Abstract*−With the upcoming deluge amount of data the number of services are emerging on internet. The retrieval of user input from web is difficult . To overcome this challenges a new approach is proposed, which is the collaborative filtering (Club-CF) and description logic based matching technique is used to solve the matching problem. Its role aim at recruiting the similar services in the same clusters of recommended services collaboratively. This approach is achieved using two stages like all the available services divided into an small-scale cluster and then the collaborative filtering algorithm is imposed on clusters. Now the number of services in clusters which is comparatively much less than the services available on web. As a result thus approach helps to reduce the online execution time of collaborative filtering. So the user recommended services were easily extracted from the database.*

Keywords: *Description logic, Collaborative filtering, Feature selection, Clustering.*

1.Introduction

In emerging trends big data which plays an vital role in the information technology. Nowadays the scales of data are quickly increased. The amount of data was also increased explosively, so analyzing the data is too difficult. Hence "big data" become a competition underpinning new waves of an productivity growth, consumer surplus and innovation. Big data finally solves the problem of integral part. In big data past few years, fetching some particular data which is too difficult. But in this project we achieved to fetch particular items in database. For this clustering based approaches were used. Map reduce[1] which plays an vital role .Nowadays, reduction of online execution time is must, so we achieved that in this project.

2. Related work

Map reduce which provides an interface for cluster computation. Here two functions were implemented by user to access the automatic distribution and functions executed by machines. First map function assign two values for input the values are axioms, and outputs values are (Key,Value) pairs [2].Then reduction function called for each key. It process all values and list result for output. To reduce work of machines assign the keys of map output. One application in map reduce[3] is ontology reasoning, it is the closure of computation of large RDFS(S)[4] graph. Thus the RDF schema rules were implemented by Map reduce jobs. For example:RDFS subclass rule

(i.e.,(1)) s rdf: type x & x rdf:subClassOf y) s rdf:type y

which predicates the key of triples \rdf :type" is the object and which predicates the key of triples \rdfs:subClassOf" is is the subject of the triple.Two Types of changes in OWL ontology's[5]:changes at terminological level(T-Box) and at assertion level(A-Box).Nowadays,the highest layer has reached maturity in ontology[6] layer in the form of description logic based languages, with OWL and DAML + OIL [7].Generic theories have to be connected to the object layer, to analyze an concrete specification[8].Nowadays, many enterprises focusing their business on internet. so it is very important to developing a new e-marketplace[9] technology.OWL_S it is the ontology web language for web services.

Pseudo code

```
Class Map per
   method Map(doc  id, doc d)
      forall term t in doc d do
         Emit(term t, count 1)

Class Combiner
   method Combine(term t, [c1, c2,...])
      sum = 0
      forall count c in [c1, c2,...] do
          sum = sum + c
      Emit(term t, count sum)

Class Reducer
   method Reduce(term t, counts [c1, c2,...])
      sum = 0
      forall count c in [c1, c2,...] do
          sum = sum + c
      Emit(term t, count sum)
```

3.Preliminaries

3.1. Challenges in massive data

For the classification, massive data was a great challenge. For dealing massive datasets map reduce programming framework plays a vital role. Also many number of machine learning algorithm available for good subset of feature selection. For efficient feature selection for example big data in healthcare, fast clustering based feature subset selection is needed.

3.2. Map reduce

The various features of map reduce are simplification, fault tolerance and scalability. This approach parallelizes large scale data processing in web indexing,data mining and bio-informatics. It is run very fast running for various applications.

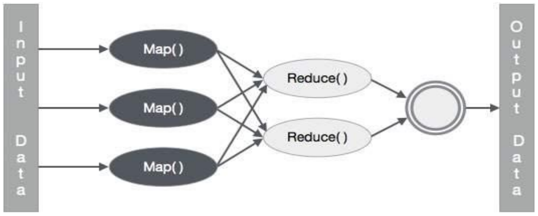


Fig.1.Mapreduce

3.3. Feature sub-selection

The process used for identifying subset is feature selection. The data contains many irrelevant features when using a feature selection technique. More information is not provided by redundant features. Feature selection defined as the subset of the more general field o feature extraction. For data analysis process, feature selection process is used. So, therefore effective way for reducing dimensionality,increasing learning accuracy and improving result comprehensibility can be done with feature sub-selection.

3.4. Generalization based clustering

Graph theoretic methods were used in many applications and studied well,which comes under cluster analysis .A simple which one is general graph-theoretic clustering. It is used to compute a neighborhood graph of instances, and any edges in the graph may delete which may be shorter or longer than its neighbors. Thus the result only each and every forest and tree become cluster. In this project, the graph-theoretic clustering was applied to the features.Particularly,the minimum spanning tree (MST) based algorithm were used,hence they do not assume the data points around centers that is used widely.

4.Reduction of online execution time: algorithm

In big data, to handle large datasets there are many algorithms were used namely:Genetic algorithm, decision tree, association rules. K anonymization algorithm and slicing algorithm.First, the genetic algorithm it is a nature inspired heuristic approach which solve the problem of search based and optimization problems. It originally developed by John Holland (1975).The basic process for a genetic algorithm is:Initialization,evaluation,selection, crossover,mutation and termination.Nowadays,many current technologies widely using machine learning.In that one of the fundamental machine learning methods is Decision tree due to its fast learning tasks and consistent prediction results.Decision tree is an decision-making technique that is commonly used to making an graphical representation of the possible consequences of an number of given cases. Association rules are if or then statements that will helps to uncover the relationships between seemingly unrelated data in a relational database. Slicing in data analysis which is used for reduction of a data in body into smaller parts.This slicing which is also compared with drilldown.It is also a process used to divide information into finer layer in hierarchy.In k anonymization algorithm, a table with n rows(records) and m columns (attributes) represented in database. The alphabet of a dataset(∑)-the range of values that individual cells in the database can take. It is represented in the symbol of ∑U{*}.A SPARQL[10] query method also used.

4.1 Association rule

An association rule is an implication expression of the form X→Y, where X&Y are disjoint item sets (i.e.) $X \cap Y = \sigma$.The strength of association rule can be measured in terms of its support and confidence.

(i.e.,(2)) Support, $s\ (X \rightarrow Y) = \sigma\ (X \cup Y)/N$

Confidence, $c\ (X \rightarrow Y) = \sigma\ (X \cup Y)/\sigma\ (X)$

5. Potential application

In bigdata,as the collection and the use of large data sets that combined and distributed to identify the patterns and create a new data based on this insights which is known as metavariables.It helps to increase the most effectiveness and the efficiency of consumer finance products.Four trends with the value for developing bigdata capabilities that support in consumer access,affordability,product quality,provider efficiency and scale.This application is applicable on the MNC Company,Big hospitalized area, and the Amazon companies.

6. Challenges in big data

The main challenges in the big data are to capturing data, to store data, to analyze data and to transfer data. In big data, for next five years there is a chance to occur some of the issues like a locality, the privacy and regulation, the project requirements and the human resources. Some of the other challenges in data visualization are: meeting the need for speed, understanding the data, addressing data quality, displaying meaningful results. The main drawback in big data is data traffic.

7. System architecture

From this system architecture, the inputs may get from laptop, internet and mobile devices. Those three storage device which need a storage space and it store in the format of .txt,.json,.doc,.pdf,.csv. But while we using hadoop, it has only one format, that is .json. Thus the .json which has two parts one is value and another one is key. That the values also stored in HDFS and it connected with web services.

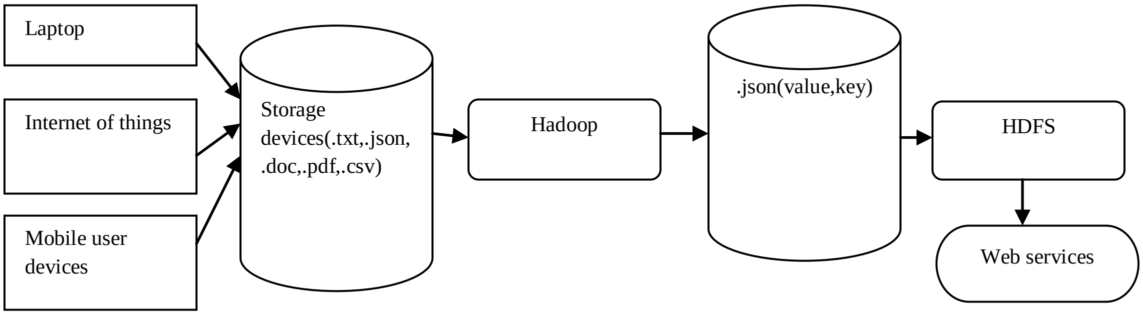


Fig.2.System Architecture

8. Proposed system

In this project,newly two techniques were used namely:Recommender system(RSs) and the Collaborative filtering(CF).Recommender system is an intelligent applications to assist the users in a decision making process in that where the people they want to pick some of the items among the overwhelming of services.Second,the collaborative filtering which has item and user based which has item and user based which is a dominant technique applied in RSs.In user based CF, it has rule of if those the people who agreed in the past, they have to agree again in future.The item-based CF algorithm recommends similar items what people have preferred before.Consequently,the service recommendation based on similar services would be either loses its timeliness.

9. Performance evaluation&10. Final Output

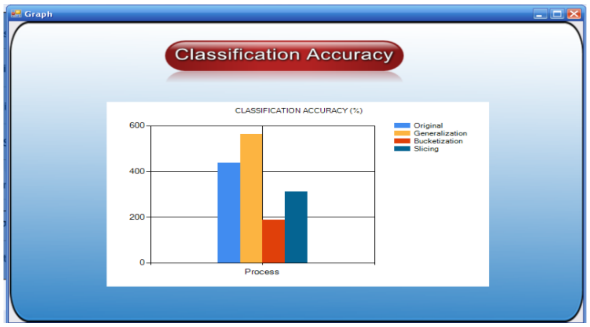
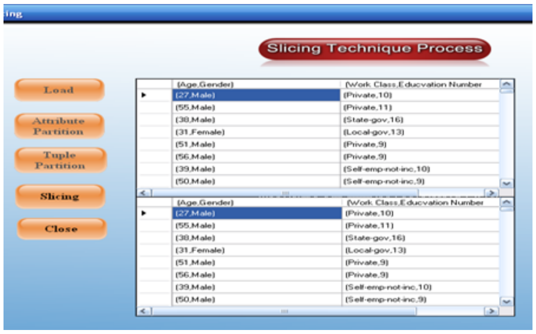


Fig.3.Performance Evaluation.



	(Age,Gender)	(Work Class,Educvation Number
▶	(27,Male)	(Private,10)
	(55,Male)	(Private,11)
	(38,Male)	(State-gov,16)
	(31,Female)	(Local-gov,13)
	(51,Male)	(Private,9)
	(56,Male)	(Private,9)
	(39,Male)	(Self-emp-not-inc,10)
	(50,Male)	(Self-emp-not-inc,9)

	(Age,Gender)	(Work Class,Educvation Number
▶	(27,Male)	(Private,10)
	(55,Male)	(Private,11)
	(38,Male)	(State-gov,16)
	(31,Female)	(Local-gov,13)
	(51,Male)	(Private,9)
	(56,Male)	(Private,9)
	(39,Male)	(Self-emp-not-inc,10)
	(50,Male)	(Self-emp-not-inc,9)

Fig.4.Sliced Data

11. Future enhancement

This project is applicable on the MNC Company in healthcare centers and the Amazon Company. The future enhancement in this paper is,it can be applied on the IOT.

12. Conclusion

This paper motivates the several directions for an future research in big data using Hdfs. In this paper, slicing achieves where each attribute is in exactly one column. By using overlapping slicing, which duplications an attribute in more than one column. So it helps to release more attribute correlations. This could provide an better data utility but privacy not assured instead need to be carefully studied and understood. In our project, random grouping which is not very effective. So we plan to design an more effective algorithm is tuple[11] grouping algorithm. Thus slicing is an promising technique to handle the high dimensional data. Hence privacy was achieved by partitioning [12] attributes into column. For example slicing applied in anonymization transaction database. Randomly generated the associations between column values of bucket, this may lose data utility. So designing data mining tasks using anonymized data which computed by using different anonymization techniques.

REFERENCES

[1] Jeffrey Dean and Sanjay Ghemawat,"Mapreduce: Simplified Data Processing on Large Clusters".

[2] A.Schicht and H.Stuckenschmidt, "*Map Resolve" in proc.5, Int conf RR Galway Ireland", pp.294-299,* (Aug 2011).

[3] Jacopo urbani, Spyros kotoulas, Eyal oren, and Frank van Harmelen,"*Scalable Distributed Reasoning Using Map reduce".*

[4] Jacopo urbani,Spyros kotoulas,Jason Maassen,Niels Drost,Frank seinstra,Frank van harmelen,"*WebPIE:a web scale parallel inference engine".*

[5] G.Antonis and A. Bikakis,"*DR-Prolog: A system for reasoning with rules and ontologies on the semantic web", IEEE Trans.Knowl.Data Eng., Vol.19, no.2, pp.233-245,(Feb.2007).*

[6] M.Jenifer, P.S.Balamurugan, T.Prince,"*Ontology Mapping for Dynamic Multiagent Environment*".

[7] Li Ding,Pranam kolari,Zhongli Ding,Sasikanth Avancha,Tim Finin,Anupam Joshi,"*Using ontologies in the semantic web:a survey",*(july 2005).

[8] K.Vijaya Kumar and G.Nanda Kumar, P.Sudha, A.kumaresan,"*Geographical approximate string search for retrieving errorious data in spatial database",*(2014).

[9] Chin-Pang, Jingzhi Guo and Zhiguo Gang, *"Inference on Heterogeneous e-marketplace activities".*

[10] Jesse Weaver and James A.Hendler,"*Parallel materialization of the finite RDFS closure for Hundreds of Millions of triples".*

[11] Hang Xiang Pan, Yingjie Li and Jeff Helflin,"*A Semantic web Knowledge base System that supports Large scale Data Integration".*

[12] Piotr szwed,"*Video-event recognition with Fuzzy Semantic Petrinets*".

[13] Jutta Eusterback GMD,Rheinstr.Darmstadt,Germany,eusterbr@darmstadt.gmd.de,"A ,"*A Multi-layer architecture for knowledge based system synthesis".*